Chapter 3: Mine Drainage Data Analysis Algorithm

A flow chart outlining the data analysis algorithm for determining baseline pollution load is shown in Figure 3.1. The algorithm includes evaluations of data quality, univariate statistical analyses, bivariate statistical analyses and time series analyses. The algorithm also includes steps to evaluate the normality of the frequency distribution and logarithmically transforms the data if the distribution is not normal (i.e., positively skewed); however, the use of the statistical methods in the algorithm does not require the distribution to be normal.

All of the statistical analyses included in the algorithm are contained in the MINITAB[1] computer software package, which was used to assess the data presented in this report. The analysis contained in MINITAB was incorporated into the REMINE[2] computer software package developed by EPA, PA DEP, and Pennsylvania State University. Other software packages included Statistical Analysis Software (SAS) and Stat Graphics. A significant feature of the algorithm and the MINITAB program in general is that a user with limited statistical analysis experience can perform the rudiments of the baseline pollution load analysis without encountering too much difficulty, while the user with greater statistical training can expand the statistical analysis to include a much greater array of statistical methods if desired. The remainder of this chapter is devoted to explaining the elements of this remining data analysis algorithm.

Data from six study sites were submitted to the standard procedures shown in Figure 3.1. (These data are described in detail in Chapters 4 through 8.) There are twelve steps in the complete analysis, and it should be emphasized that only the first nine are needed for routine remining permits. Steps 11 and 12 are for research purposes only. The most important step is initial examination of the data (Step 1, Figure 3.1). Following this examination, missing values are identified and adjusted. Additionally, any extreme outliers (Step 3) should be examined to see if they are real observations or errors of entry at some stage in the data collection procedure.

The next step (Step 4) is to graph discharge (flow) versus days (ordered observations). Frequently, it is advisable to plot log discharge in order to reduce extreme variations. This procedure also helps to reduce extreme positive skewness (if present in the data). This reduction of asymmetry improves the subsequent analysis of the data and makes the probability statements more reliable. Because extreme observations may result from unusual events (such as heavy downpours, snowmelt), the reduction of variation should be used with discretion. In many cases, these unusual extreme data values may indicate events of considerable importance in the study of the natural variation in the data.

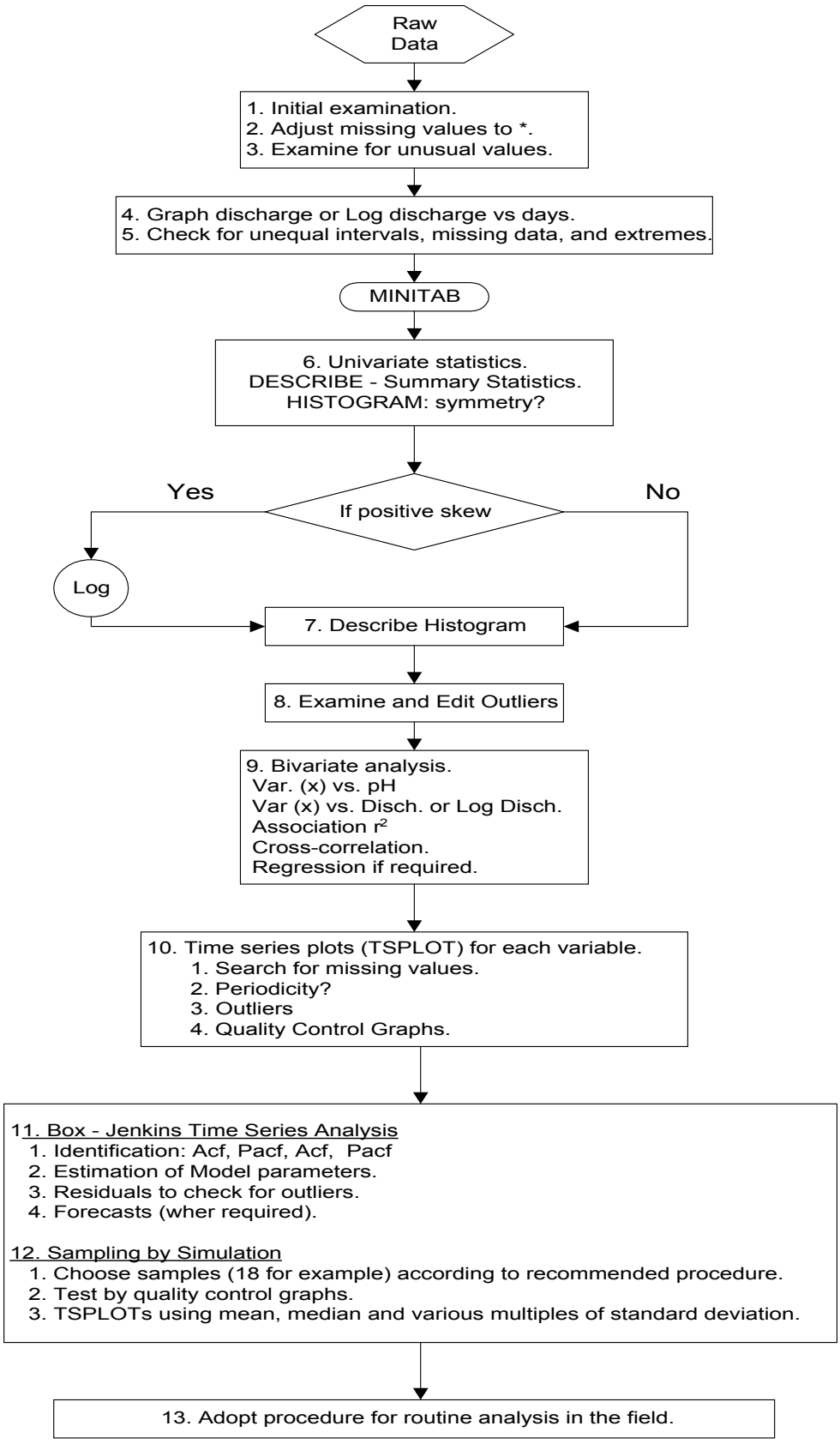
---

[1]MINITAB is a commercial software package from Minitab, Inc. ©1986, 3081 Enterprise Drive, State College, PA 16801.

[2]REMINE is a computer software package developed by EPA, PA DEP and the Pennsylvania State University, Version 1.0 (November 1988) and Version 2.0 (April 1992), page R-2.

Step 5 is also crucial for determining regularity of the sampling to further identify the larger gaps in the data. The plot of discharge versus days prepared in Step 4 is one way of seeing this aspect of behavior in the data. Another useful procedure is to take "first differences" of days (or order of observation). This procedure leads to a frequency distribution and a histogram in which the intervals between observations are clearly displayed.

Figure 3.1 Algorithm for Analysis of Mine Drainage Discharge Data



Ideally, all the intervals between the observations should be equal; in practice, this is rarely achieved. One or two days on either side of the ideal date is adequate for fourteen-day intervals. Many gaps of five or six days make the subsequent analysis much less exact and larger intervals (e.g., 90 days) make the analysis more difficult to interpret correctly. Large gaps in the data preclude rigorous time series analysis which requires a very close approximation to equal intervals between observations. In general, the more sophisticated the statistical analysis, the more sensitive it is to data gaps.

As a general recommendation, it is helpful to insert a missing data symbol (e.g., *) where there are data gaps (i.e., a few missing flow measurements or a few missing values for water quality parameters) and produce the mean, median, standard deviation, etc. of the truncated data set. If the frequency distribution of the variable is reasonably representative (e.g., symmetric), or has been made so by log transformation, then the means may be substituted for each missing data symbol (*) and the frequency distribution and summary statistics (mean, median, standard deviation) rerun on more complete data. Of course, insertion of the mean does not gain information; it only makes subsequent analysis more correct. If the data are asymmetric, the median is a more representative estimate of the "central tendency" and should be used rather than the mean.

This entire procedure (Steps 1–5, Figure 3.1) is aimed at "massaging" the data into a form suitable for statistical analysis. If there are only a few observations (18 or so, for example) it is somewhat arbitrary whether or not one wishes to smooth the data, because very little extended analysis will be appropriate.

Univariate Analysis (Algorithm Steps 6, 7, and 8)

In Step 6, the data are analyzed and plotted to obtain the summary statistics and to examine graphical displays of the data to determine the presence of skewness and extreme values. Stem and Leaf plots can be used in place of histograms of frequency distributions as shown in Figure 2.4.

This procedure includes calculating statistics for each individual variable (univariate statistics). An example of this procedure is displayed in Table 3.1 of the summary statistics for the analysis of the data from the Clarion site (discussed in Chapter 5). In this example, there are seven parameters and eleven summary statistics that were calculated using the REMINE program. There are N = 96 observations (column 1 of Table 3.1); N* (column 2) is the number of missing observations (19 for the discharge variable). Columns 3 and 4 list the means and medians respectively. Column 5 is a special kind of mean, called by Tukey (1977, p. 46) the "trimmed mean." Columns 6 and 7 contain the standard deviation (STDEV) and standard errors (SEMEAN) of the mean as measures of spread. Columns 8 and 9 list two extremes (min and max) yielding the range of the values. Columns 10 and 11 contain the quartiles (Q_1 and Q_3), yielding a measure of spread around the central tendency (mean or median); this spread is less sensitive to the extremes and so is often preferred in distributions which are irregular (e.g., strongly skewed). The coefficient of variation (Column 12), usually expressed in percent (CV%), is defined as the ratio of the standard deviation to the mean multiplied by 100. This is a

useful approximate guide to the degree of variation in a parameter. In general, a CV < 30% represents a stable (in control) parameter. Most of these parameters, however, show much larger variation, principally because of the large effects of extreme events.

Table 3.1: Summary Statistics for S3CLAR (N=96)

	N	N*	Mean	Median	Trimmed Mean	Standard Deviation	Standard Error of the Mean
pH	96	0	3.696	3.195	3.612	0.985	0.101
Discharge	77	19	12.58	6.30	9.00	22.66	2.58
Acidity	96	0	522.4	483.5	505.6	346.4	35.4
Total Iron	96	0	82.40	75.00	79.31	51.01	5.21
Ferrous Iron	96	0	54.84	39.50	47.44	66.99	6.84
SO_4	96	0	1528.	1569.0	1525.9	566.0	57.8
Ferric Iron	96	0	27.56	23.60	31.01	70.58	7.2

	Minimum	Maximum	First Quartile	Third Quartile	Coefficient of Variation
pH	2.670	6.430	3.002	4.455	26.6
Discharge	0.05	172.00	3.59	12.54	188.1
Acidity	1.0	1546.0	232.5	737.7	66.3
Total Iron	8.70	257.00	39.70	110.25	61.9
Ferrous Iron	0.90	612.18	25.12	68.60	122.2
SO_4	296.0	3241.0	1181.5	1878.2	37.0
Ferric Iron	-581.68	152.00	5.00	55.00	256.1

A second series of statistics referred to as letter values (e.g., H-spread) is sometimes calculated to identify various measures of spread. These spreads can be used to set limits for water quality (see Tables 8.6 and 8.7, (Q_3-Q_1)). These letter values (LVALS) were first defined by Tukey (1977, p. 22) and are mentioned in the MINITAB Reference Manual (p. 168). These values are best described in Velleman and Hoaglin (1981, p. 33).

If the data are positively skewed (i.e., skewed towards the high end of the values on the variable scale) the data should be logarithmically transformed and the univariate analysis repeated (Step 7, Figure 3.1). The log transformation tends to make the histogram more symmetrical, although there is a tendency to over-correct in some cases and introduce negative skewness.

It is possible to use another transformation such as the square root of the variable, which may well suffice to avoid over-correction that came from the logarithmic change. The use of various transformations is reviewed in Tukey (1977, Chapter 3) and specifically for symmetry, in the MINITAB Handbook (p. 72 – 76) and the MINITAB Reference Manual (p. 50 – 52). It is also discussed in Velleman and Hoaglin (1981, p. 46 – 49) and Box and Cox, (1964).

In evaluating the statistics produced for the transformed data, the user should be cautious of the coefficient of variation values. Use of the coefficient of variation with log transformed data may result in extreme distortion because the transformation leads to a mean of small value. This results in a denominator of the ratio that is small resulting in a CV that is inflated.

Step 8 in Figure 3.1 is used to check and accept or modify outliers. Outliers tend to inflate the variance or spread of the data and make the statistical tests less sensitive. For this reason, outliers should be reduced only after deciding that such extreme values are not "real" or when it is specifically desired to make the statistical testing more sensitive. As mentioned earlier, some outliers are indicators of unusual events (e.g., floods, storms) and should not be removed or even subdued, but instead should be used to reflect the occasional unusual events.

Bivariate Analysis

The next step in data analysis (see Step 9) concerns the relationship between pairs of variables (bivariate analysis). If two variables are closely associated (e.g., a correlation coefficient, $r > 0.8$), both may be reflecting the same source of variation and one may be considered redundant. It is possible to use this kind of feature to select the simpler test (or less expensive analyte) and ignore the other parameter in subsequent studies. Sometimes several variables reflect the effects of the same events.

One expects, for example, pH to decline with increasing acidity and sulfate. In the case of calcium and manganese, on the other hand, one expects sympathetic variation. If examination of the data shows that this expected relationship is not present, the reason for its absence should be sought.

The correlation coefficient (r) is usually used to represent the (linear) relationship between any pair of variables. The coefficient of determination (r^2) is, however, a better measure of the intensity of the association between a pair of variables; for example, $r = 0.7$ looks large because the range of r is from –1 to +1, but it means that $r^2 = 0.49$ or 49% of the variation is common to the two variables and there is 51% of the variation "unexplained" by the association. It is necessary, therefore, to realize that one needs $r > 0.8$ to claim that a strong association exists (i.e., > 64% in common).

Another feature which is illuminated by using r^2 as well as r is the statistical test which accompanies a specific value of r. For a sample size of N = 174 (Table 6.3), a value of $r > 0.124$ is significantly different from zero at the five percent probability level. This should be accompanied by the corresponding value of r^2. In Table 6.3, the correlation coefficient between pH and acidity is $r = -0.365$. This value comfortably exceeds the $r = (\pm) 0.124$, thus it is statistically significant. Nevertheless, the corresponding $r^2 = 0.133$ means that only 13.3% of the variation is common to both variables.

In the graphs presented in Figures 6.5a and 6.5c, the variation of both parameters increases as their values increase. This phenomenon is called heteroscedasticity. In general, it is advisable to plot the logs of the variables which tends to make the variables homoscedastic. Since

heteroscedastic variables show a difference in variability with changes in values of the parameter, no probability statement should be made without transformation so that the variables are homoscedastic. Peculiarly, the change from heteroscedasticity to homoscedasticity does not lead to a major change in the value of r, but does make the probability statements more reliable.

One more avenue should be explored in bivariate analysis, and that is to determine whether there is any lag in correlations between pairs of variables. Cross-correlation analysis is performed to see if a weak relationship at zero lag may be much stronger at greater lags. This could result from a delayed effect. For example, suppose discharge increases and sometime later, pH drops. Correlation at zero lag may be quite low, but at some higher lag it may increase showing that it takes time for the effect of changes in discharge to affect pH or some other variable. The cross-correlation function (CCF) is the measure used for this purpose. For example, suppose that an event occurs and affects one variable immediately but only affects another variable five observations later. In this case, the linear correlation coefficient at zero lag may be quite low but may show a strong association after a five day lag. The cross-correlation function calculates the linear association between observation 0 to t days apart and so gives a picture of when the association is strongest. The range of t is from – (sqrt (N) + 10) to (sqrt (N) + 10) where N is the number of observations in the series. In most of the examples presented in this report, there did not appear to be any lag in the effects.

Time Series Analysis

The remaining steps (10 through 12, Figure 3.1) were used to set up baseline behavior based on relatively long data records. In this way, expected behavior of various parameters are established for comparison with the shorter data records that are commonly used in routine remining permitting. The likelihood of unusual events is then displayed, and the frequency of a single or a few unusual observations may be used to judge how often these events occur. In this way, these events can be distinguished from other departures that lead to warnings, triggers, or exceedances in pollution load and therefore, would be less likely to result in false alarms.

One procedure which is readily available as part of the full Box-Jenkins treatment, but was not used in these studies, was Transfer Function analysis. This analysis would be a most attractive way to correct variation in some parameter (e.g., Fe) for variation in flow and then proceed to analyze the residual variation in the parameter after the effects of flow were removed. This would also be an alternative way of looking at the "load" variable in place of concentration.

Similarly, there is a procedure in Box-Jenkins analysis called "intervention analysis" which may be used to compare and contrast variation in a parameter before and after treatment is applied. This has obvious applications to remining operations. Needless to say, use of these procedures requires an extensive set of observations taken at equal intervals, with few data gaps.

Variation in many of the parameters, from the different locations, appears to follow a common pattern. There is usually some type of gradient present in the data which may be increasing or decreasing over time. This results in a typical autocorrelation function (Acf) pattern and a large spike at lag 1 in the partial autocorrelation function (Pacf). This trend should be removed before

fitting a model. This is best done in nearly all the examples in this particular series of investigations by taking first differences of the variable of interest. The subsequent model-fitting usually leads to a moving average model. In Box-Jenkins notation this is an IMA (0,1,1) model. It is essentially a random walk after first differences are taken.

Quality Control (QC) Limits

Step 10 of the algorithm on times series plots of the variables (Figure 3.1) includes an item (# 4) on quality control graphs. Items 2 and 3 of Step 12 (sampling by simulation) also refer to quality control graphs. The final step of the algorithm (Step 13) is a procedure for routine statistical analysis of data contained in remining permits. From the discussion of quality control throughout Chapter 2, it is obvious that the development of a relatively simple quality control approach for mine drainage data analysis is a major objective of this report and a significant component of the routine procedure in Step 13 of the algorithm. Chapters 4 through 8 contain further discussion, tables and plots of various examples of quality control limits. Examples from the six mine drainage case study sites lead to the statistical summary and review of quality control limits in Chapter 9, wherein options for the routine use of quality control limits are presented.

Throughout this report the conventional quality control limits based upon the mean and standard deviation of the normal frequency distribution are compared to another set of non-parametric quality control limits based upon the median and other order statistics (e.g., quartiles, H-spreads, C-spreads), which may be more applicable to mine drainage data that frequently do not conform to normality. The quality control options in Chapter 9 of this report are a component of the routine procedures for establishing baseline pollution load and monitoring in remining permits. These procedures are related to the recommended statistical procedures set forth in Chapter 3 and Appendix A of EPA's *Coal Remining Statistical Support Document*. However, the user of these routine procedures should be ever mindful that no single set of quality control limits or specific statistical test will be perfectly applicable to all mine drainage sets or even to all discharge parameters within the same data set. The user should carefully examine the data and follow the fundamental steps of the algorithm in order to properly use the statistical tools that are most applicable to the characteristics of the data.

In the following chapters, more than one equation was used to calculate QC interval spreads. These equations were chosen based on the distributions of the parameters collected in the given data sets (i.e., number of results, amount of variability, lack of normality, etc.)

The first equation ($\overline{X} \pm 2\hat{\sigma}$) is based on the typical confidence interval for a mean under the normal distribution. However, unlike the typical equation for a confidence interval around a mean, the standard deviation was not divided by the square root of the number of results (N). The exact interpretation of the usual confidence interval is that the true mean of all post-remining results for the given site will fall into the calculated interval with 0.95 probability. For the purpose of quality control, however, this interval may be extremely tight, given the large number of results collected for each dataset. For baseline permit pollution load data sets, the number of results collected would likely be much less, and therefore would produce wider

intervals. A different value of N (N') could be used in the equation, reflecting the number of results likely to be collected and used to calculate the mean that will be compared to the interval. For example, if monthly samples are collected for a year, N' would equal 12. However, if the purpose of the interval is to evaluate individual results rather than a mean, then N' should equal 1. This is what was done in Chapters 4, 6, 7 and 8, where the above equation is used.

The two other equations that are used in quality control tables in the following chapters are non-parametric, in that they do not require that the collected data follow a normal distribution. They are based on the non-parametric equivalent of the mean (the median) and the non-parametric statistic for variability (the interquartile range). The first interval,

$$\text{Md} \pm 1.96 * (\frac{1.25 \ * \ \text{H} - \text{spr.}}{1.35 * \sqrt{N'}}),$$

is discussed in McGill, Tukey and Larsen (1978). This interval is used to assess whether a median follows the same population as the baseline pollution load data, and is therefore divided by the square root of N', where N' is the expected number of remining results. The chosen multiplier, 1.96, is appropriate when it is assumed that the variability of the baseline data and the remining data are approximately equal. However, if the variability of the baseline data and remining data are different, a smaller multiplier (1.39) is appropriate. When it is not known whether the two variances will differ, the midpoint of 1.39 and 1.96, (i.e., 1.7) could be used. The above equation is used in Chapters 6, 8 and 9.

A second equation, Md ± 1.58 * (H-spr.), was used in Chapter 4. In this second equation, the value of 1.58 was chosen by using the midpoint multiplier (1.7) and simplifying the equation by multiplying by 1.25 and dividing by 1.35. The purpose of this equation differs from the previous one, in that it is designed to evaluate individual results, rather than the remining median.